Evaluation of Human-robot Interaction in the NIST Reference Search and Rescue Test Arenas

Jean Scholtz, Brian Antonishek, and Jeff Young
National Institute of Standards and Technology

Abstract

In this paper we discuss studies we have been conducting on human-robot interaction (HRI) during the Urban Search and Rescue (USAR) competitions in the NIST Reference Test Arena. We discuss some of the analyses we have already done on the data we have collected and present the guidelines we have produced based on these studies. We discuss future plans for augmenting USAR competitions to specifically compare different methods of HRI.

Introduction

The ultimate evaluation of how humans and robots interact is the measure of their combined performance. In search and rescue the human-robot team has two goals: to locate victims and to provide accurate information about their location and their alertness state to human rescuers. These goals need to be achieved under a number of constraints. Teams need to operate for extended periods of time; the number of personnel used in the operation should be limited due to the dangerous nature of the operation; and the tasks need to be accomplished quickly to maximize the lives that can be saved [2]. Many human-robot search and rescue teams have participated in Urban Search and Rescue (USAR) competitions in the NIST test arenas [7,8]. The overall scoring for these competitions emphasizes these goals and constraints. Although scoring varies from year to year, the teams are rewarded for locating victims in a timely fashion, accurately assessing their condition, and providing good maps for rescue workers. Teams are penalized for causing further damage to the collapsed structure. Teams requiring multiple operators for individual robots are also penalized.

Good human-robot interaction (HRI) contributes heavily to a team's overall score. However, there are a number of other contributing factors as well, including the mobility of the robot, the skill of the operator, the robustness of the hardware, software, and communications, and the sensory perception provided. We are interested in evaluating the various user interfaces to determine what information and information presentation contributes to the overall performance of the system.

Pros and Cons of Using the USAR Competitions for HRI Evaluation

The primary benefit of using these competitions for studying HRI evaluation is that the competitors provide many more ideas for user interfaces than we, as researchers, could possibly prototype and test.

The limitations are that we can only study the operator role [11]. The operators in the competitions are expert users, i.e., robotics researchers. We are not allowed to interfere with the competition environment which means that we cannot collect think-aloud or talk-aloud protocols [5,6] from the operators. It is difficult to interview the operators after their runs as they are busy getting ready for their next round. Moreover, the teams come from allover the world and there are language barriers to overcome. The user interfaces and the robots are dynamic. The teams make changes during the competitions. Different robots are used; different sensors are used; different teammates take turns at being the operator.

In spite of the limitations, these competitions provide a rich source of data in a reasonable USAR simulation.

Data Collection

We have collected data at six major competitions starting in 2002. We collect video data of the user interface, the operator, and the robot as it moves through the arena. In addition we collect information about the robot's path and coverage of the arenas. We also have access to the overall performance scores including penalties occurred.

We typically tap into the video output of the operator control unit (OCU) and direct this to a scan converter which sends the converted output to a video recorder for later analysis. As the setup time for teams to get ready for their rounds is between 10 and 15 minutes, data collection setup has to be quick and flawless. Prior to the initial rounds, we test out the data collection equipment with each team who agrees to participate in our study. We make sure that all the video is time stamped so that we can easily move between the operator view of the user interface and ground truth as represented by the robot moving in the arena during analysis. It is difficult to tape the movement of the robot in the arena, as portions of the arenas are covered. Debris and multiple levels in the arenas make it difficult to see the robot at all times without being physically in the arena. We try to capture data from outside of the arena as our presence can cause the sensors on the robot to mistakenly identify us as victims or unintentionally point out possible paths through the debris. Figure 1 shows three different sections of the Robocup 2004 NIST test arenas.

Figure 1: NIST Test Arenas at the Robocup USAR 2004 Competition

Analysis of Data

We have completed analysis of two sets of data at this point in time. Our initial data analysis was completed on data collected at the 2002 USAR competition at the American Association of Artificial Intelligence [14]. We collected data from all the teams in the competitions but we coded only data from the four top ranking teams. We used the data from the semifinals and finals. We were interested in looking at how the overall performance correlated with a finer analysis of performance. We looked at the video tapes and coded the amount of time each team spent in navigation or monitoring navigation, in identifying victims, in logistics, and in failures. Table 1 contains the definitions of these terms.

Table 1: Definitions of Coded Activities

Activity Coded	**Definition**
Navigation or monitoring navigation	This activity was coded when operators were teleoperating a robot, or in the case of semi-autonomous robots, when the operator was issuing navigation commands and watching the user interface to assess how the robot was moving.
Victim identification	We coded this activity when the operator thought he had sensed a victim and moved closer or used other sensors to assess the status of the victim.
Logistics	Activities such as starting up another robot were coded as logistics.
Failures	Hardware, software, and communications dropouts were coded as failures.

Table 2 shows the percentage of times the four teams spent in these activities. Note that we were only able to code two of the three runs due to issues with the data collection mechanism. The total time is given in minutes. Each team was allocated 15 minutes for their runs. It was difficult to actually coordinate with the competition officials to know the actual start and stop times and, in one case, we lost some time due to a data collection issue. Note that the percentages do always add up to 100%. This is basically due to rounding areas in calculating times.

Table 2: How teams spent their time

	Run	**Total Time (min)**	**% Time**			
			Navigation/ Monitoring Navigation	**Victim ID**	**Failure**	**Logistics**
Team A	1	10:39	46	51	0	3
	3	14:45	62	18	19	1
Team B	1	14:33	81	19	0	0
	3	16:42	77	23	0	0
Team C	1	13:26	59	23	17	0
	3	14:39	69	12	18	0
Team D	1	15:12	55	32	0	12
	3	13:30	87	4	0	9

We found that teams using some sort of automatic mapping were more successful in navigating the arenas. Operators who had to keep maps in their heads became confused about where they were at times. We looked at the penalties incurred by the teams and found instances where the operators were unaware of the surroundings of the robots or the status of the robots. In particular, few robots in this particular competition had a view of what was behind them. In situations where the operator was forced to back up or to make a series of tight turns this resulted in penalties for bumping into walls or victims.

In our analysis of a second set of data collected at Robocup 2003, we looked at issues of awareness [9]. Burke [1] identified situation awareness (SA) as a major component needed for effective human-robot performance. Scholtz[10] has modified Endsley's SAGAT [4] methodology for measuring the SA provided by supervisory interfaces for semi-automonous driving vehicles. Scholtz also analyzed the time needed for operator acquisition of SA in two types of terrains [12, 13]. In this analysis we used a modification of awareness tailored to HRI [3].

If we consider teams consisting of humans and robots, we can define 5 types of awareness:

- Human-robot awareness
- Robot – human awareness
- Human-human awareness
- Robot – robot awareness
- Humans' overall mission awareness

In the majority of teams competing in the USAR Test Arenas, we are able to evaluate only human-robot awareness. That is, does the human have knowledge of the location, status, and behavior of the robot? We find few teams that have multiple robots with any collaboration capabilities (robot-robot awareness) or use multiple operators (human-human awareness). Moreover, the current generation of robots in these competitions has no awareness of the operators' status (robot-human awareness).

We used an indirect means of assessing human-robot awareness as we are not able to intervene and ask the operator to verbally describe any given situation. We coded critical incidents observed in the video tapes of the robot moving in the arena. Critical incidents are defined as a situation where the robot was in a position that could potentially be harmful to the robot, the environment, a victim, or the mission. Originally, we had intended to code critical incidents that were "avoided", such as when the robot was able to move through an extremely tight space without causing any damage. However, we found that we were unable to do this consistently. We were able to consistently locate and code critical incidents that had a negative outcome, e.g. the robot bumped into a wall. We classified the critical awareness incidents into one of five categories: global navigation, local navigation, victim identification, obstacle extraction, vehicle state. Obstacle encounter was coded when the robot had actually run into an obstacle and had to perform maneuvers to free itself. Vehicle state awareness was coded when the operator did not realize that the robot was in other than a normal state, e.g. tipped over. In the runs we coded, we found evidence of critical incidents only in the categories of local navigation, obstacle extraction, and vehicle state. We did see evidence of the other types of critical incidents but these were not in the actual runs selected for coding (the semifinals and finals). Table 3 shows the numbers of critical incidents occurring for the three teams analyzed. These three teams were selected for analysis as they placed in the final round of the competition.

Table 3: Analysis of Critical Incidents by Team

	Local navigation	Obstacle encounter	Vehicle state
Team A	4	6	5
Team B	1	9	2
Team C	10	11	5
Total	15	26	12

Obstacle encounters were the most prevalent types of critical incidents. Robots became entangled in loose debris in the arenas and it was difficult for the operators to know that.

In the most recent competition, Robocup 2004, we noted that teams typically had one of two sources of situation awareness information implemented. A number of teams used some sort of overhead cameras to provide a frame of reference for the robot in relationship to the environment. Other teams had implemented some sort of automatic mapping software, using a variety of sensors, including sonar and ladar. At this point we have not had time to do a full analysis, but an early analysis looks at the five teams who were in the final runs. Table 4 shows the penalties by team.

Table 4: Penalties by type of situation awareness

Penalties for teams using automatic mapping		Penalties for teams using overhead cameras	
Team A	0	Team P	80
Team D	5	Team S2	5
Team S1	40	(note Team S2 had only 3 runs completed as they had to end one run prematurely due to a problem with the robot)	

These penalties are all local navigation penalties. That is, the robot either bumped into the walls of the arena or into a victim. While these results should be viewed as very preliminary, our impression is that the automatic mapping is more helpful in providing situation awareness. This is not surprising, as the video information, while helpful, still requires considerable interpretation by the operator. Also, if there happens to be any sort of communication interference, the video is extremely difficult to view. The teams that we analyzed were the top scoring teams, which implies that they had reasonable coverage of the area and located a number of victims. Low scoring teams may have few penalties due primarily to an inability to move very far into the USAR arena.

The majority of teams we have analyzed have been teleoperated, using autonomy only for such things as mapping. While we have had some fully autonomous teams in the competitions, they have not been successful in navigating the difficult environment in the test arena. In our first analysis, two of the teams operated in semi-autonomous modes. The operators were responsible for overall navigation, but left the local navigation (obstacle avoidance, waypoint navigation) to the robots in many instances. We intend to analyze future teams to determine how critical incidents change based on the level of autonomy.

Discussion

Based on the analyses we have completed to date, we have been able to provide some guidelines for human-robot interaction design. These are summarized below.

- Information for effective situation awareness should include:
    - a frame of reference to determine the position of the robot relative to the surrounding environment
    - indicators of vehicle state, such as pitch, roll, traction indicators, indicators of sensor status, and camera positions relative to the robot body.
    - a map to provide global navigation information
- Minimize the number of windows provided to the operator.
- Provide a fused view of sensor information.
- Support multiple robot operators in a single display.
- Provide help from the robot in determining what mode of autonomy is most useful.

To date, we have been able to analyze data collected during USAR competitions to provide some guidelines for the design of effective user interfaces for USAR robots. We are encouraged that our work is making a difference as the situation awareness offered in the user interfaces deployed in current competitions is certainly increasing. The downside of our work is that the analysis takes considerable time and by and large the results are consumed by human-computer interaction researchers, not robotics researchers.

Future Plans

We are interested in providing feedback about HRI designs in a more timely fashion and to the robotics community more directly. In the final rounds of Robocup 2004, robots were placed in an internal spot in one of the arenas. The operator had to first assess where the robot was and then devise a strategy for moving out into the arena to locate victims. We are currently working on devising extensions to this, similar to the compulsories in figure skating competitions. This would help us assess during the competition how well the operator is able to gain situation awareness based on the user interface. While the NIST Reference Arenas provide a standard area in which the robots have to perform, there is no guarantee that robots encounter the same obstacles. Moreover, due to variations in size and mobility, we cannot expect robots to do equally well in navigating the same environment.

It is important for effective search and rescue that teams are in control. This means having good SA at all time about where team members, including robots, are and what they are doing. Situation awareness could be demonstrated by placing robots in specific situations (such as close to obstacles, on different types of surfaces and grades, or near negative obstacles) and measuring the time and accuracy of the SA by the operator –robot team. We are working on a user interface design for our own robotics platform. It would be possible to consider our performance as a baseline that the teams should try to best.

References

1. Burke, J., R. Murphy, and M. Coovert. (2004). Moonlight in Miami: An Ethnographic Study of Human-Robot Interaction in the Context of an Urban Search and Rescue Disaster Response Training Exercise. *Human Computer Interaction.*

2. Casper, J. (2002). *Human-Robot Interactions During the Robot-Assisted Urban Search and Rescue Response at the World Trade Center*. MS Thesis, University of South Florida Department of Computer Science and Engineering.

3. Drury, J. L., J. Scholtz, and H. A. Yanco. Awareness in human-robot interactions. *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, Washington, DC, October 2003.

4. Endsley, M. 2000. Theoretical Underpinning of Situation Awareness: Critical Review (2000) in Mica R. Endsley and Daniel J. Garland (Eds.) *Situation Awareness Analysis and Measurement.* Lawrence Erlbaum Associates, Mahwah, New Jersey. 3-32.

5. Ericsson, K. A. and H. A. Simon (1980). "Verbal reports as data." *Psychological Review*, Vol. 87, 215 - 251.

6. Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: The MIT Press.

7. Jacoff, A., E. Messina, and J. Evans (2001). "A reference test course for autonomous mobile robots." In *Proceedings of the SPIE-AeroSense Conference*, Orlando, FL, April.

8. Jacoff, A., E. Messina, and J. Evans (2000). "A standard test course for urban search and rescue robots." In *Proceedings of the Performance Metrics for Intelligent Systems Workshop*, August.

9. Scholtz, J., J. Young, J. Drury, and H. Yanco. (2004). "Evaluation of Human-Robot Interaction Awareness in Search and Rescue." ICRA 2004.

10. Scholtz, J., B. Antonishek, and J. Young. (2004). Evaluating human-robot interfaces: development of a situational awareness assessment methodology. *Hawaii International Conference on System Science 37 (HICSS 37)*. Hawaii, Jan. 2004.

11. Scholtz, J. (2003). Theory and Evaluation of Human Robot Interactions, *Hawaii International Conference on System Science 36 (HICSS 36)*. Hawaii, Jan., 2003.

12. Scholtz, J., B. Antonishek, and J. Young. (2003). Evaluation of Operator Interventions in Autonomous Off-Road Driving. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems, (PERMIS).* Gaithersburg, MD Sept. 2003.

13. Scholtz, J., B. Antonishek, and J. Young (2004). Operator Interventions in Autonomous Off-road Driving: Effects of Terrain. To appear in *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, The Hague, NL October 2004.

14. Yanco, H., J.L. Drury, and J. Scholtz. (2004). Beyond Usability Evaluation: Analysis of Human-Robot Interaction at a Major Robotics Competition.. *Human-Computer Interaction, 2004.*